www.pipelinepub.com
Volume 22, Issue 3

AI Traffic Transforming Networks

By: Mark Cummings, Ph.D.

Generally, we think of transformation as a process triggered by a human. That is, people think about fundamentally new ways to solve a problem. Then, implement it. Today, our networks are being transformed not by human choice, but rather by evolving AI traffic patterns. This transformation is not a single-event response. Rather, to be successful, network leaders have to anticipate and plan for a series of changes in traffic patterns triggered by developments in GenAI and supporting hardware.

WANs

Over the last 60 years, Wide Area Networks have transformed from point-to-point analog, to mesh TCP/IP, to data center TCP/IP, to large numbers of data centers with local points of presence. These transformations started out with what we call forklift migrations. That is physically rip and replace. The invention of SDR (Software Defined Radio) and SDN (Software Defined Networking) helped make the transformations less costly and disruptive. But, still, changing fundamental network infrastructure architecture can be very expensive. The emergence of GenAI is creating another transformation, and network leaders need to manage the infrastructure evolution carefully to avoid costly difficulties.

Right now, a lot of planning, investment, and implementation is focused on developing a relatively small number of extremely large AI-focused data centers. These data centers are, and will become more so, magnets for traffic. So much traffic that the previous WAN structure will have trouble handling the traffic.

The WAN traffic problem is not just the amount of AI traffic, but rather the geographical concentration. That is, the relatively small number of AI data centers. This forces the construction of new physical networking resources in the same area as the AI data center and in the areas feeding into it. The investments required to do this can be quite large.

AI data center concentration is a function of the business models of the providers, the technology, and the applications (the types of uses the technology is put to). In the technology, training requires very large and growing computing resources all in a single data center. As each generation of LLM (Large Language Model, that is, the engine of GenAI) is 10X larger than the previous generation, the data center required to train it goes up accordingly. Training does not produce extremely large WAN traffic volumes. The Training corpus can be quite large. But it is moved once into the data center, where training occurs.

Individual inference requests (the process of getting output from an LLM) do not require anything near the same level of resources. But when many simultaneous requests are being served, the resource requirement goes up accordingly. In data center implementations, this is producing very large amounts of traffic.

The first business model focused on chat applications. That is, many users subscribe to a service where they can ask questions and carry on conversations with a particular GenAI system. These service providers offering inference deliver service continuously. If a data center AI is not available every once in a while, it is an annoyance. Not a crisis.

More recently, intelligent agents have been created using GenAI. Some of these agents have time-critical 24/7 responsibilities. When there was a network problem recently and data center AI's were not available, this caused crisis-level problems for some of these agents.

Training is done more intermittently. From a business perspective, it is valuable to be able to use large training resources for other purposes. Some rent out units of resource (such as Nvidia processors) on a per-minute basis. Others provide inference services via not actively employed training facilities.

Large corporations are building their own AI data centers. These are set up to run models that are configured for their specific needs. This may mean custom LLMs or additional customized training of existing LLMs. It often involves very sensitive proprietary information and proprietary processing.

This kind of concentrated AI traffic coming into a relatively small number of very large AI data centers tends to produce a hub and spoke network architecture. This is similar to what the San Francisco financial district commuter traffic produced in the San Francisco Bay Area public transportation system. That is, a transportation network architecture dominated by the requirement to deliver very large numbers to and from a center.

Emerging Edge AI

While these large data centers with hub and spoke networks are being developed to meet the ever-growing traffic demand, Edge computing capabilities are increasing. The ability to run the largest LLMs on commodity hardware with SSD streaming is here, although with increased latency. Recently, that has been enhanced, reducing latency by allowing several computers to cooperate in running a very large model. At the same time, systems are coming to market that further reduce the Edge processing limitation. Examples include the M4 (MacBook and Mini) Pro, M4 Max, and M4 Ultra series by Apple. Soon to be followed by the M5 series, with the M6 series is approximately a year away. Although Apple appears to be a leader in the Edge hardware race at this time, others are sure to rise up to challenge Apple.

Operation at the Edge has some intrinsic advantages that include: reliability, privacy/IP protection, and network latency. There may also be financial drivers as well.

The recent Amazon outage is a good example of what can happen when people or organizations depend on data center network-accessible AI. Having an Edge implementation, either as a standalone or as a backup, can overcome these outage problems.

Recently Gartner extolled Edge's advantages. "Edge computing … is evolving from a buzzword to a necessity. By processing data closer to its source, edge reduces bandwidth needs and enables instant insights … critical for IoT-heavy industries facing 5G proliferation … enhancing resilience against outages. For example, autonomous vehicles rely on edge for split-second decisions …" Latency can be important. Especially for intelligent agents, time can be critical.

Just the round-trip network communication time to and back from the data center may be problematic.

Working with vendor-provided data center AI has some inherent privacy and IP (Intellectual Property) exposures. For some applications, these exposures can be quite important. For them, the fact that their data can be used in training LLMs, or get into the context windows of other users, etc., may be too great a concern. The data may not go to others in exactly the form received by the data center. But it may be used in training. Thus, it is part of the reasoning data that the GenAI system uses. Resulting in what is termed 'IP Leakage'.

Some organizations may meet this concern by implementing their own private data center. However, this still has a data exposure risk on the network that accesses the data center. Plus, Edge systems may be more manageable, more cost-effective, have more predictable expense profiles, etc. For these reasons, and possibly just convenience, users may prefer running GenAI locally on edge systems.

Edge AI generates fundamentally different traffic patterns than Data Center AI. Instead of being hub and spoke, point to multi-point networks, Edge tends to be multi-point to multi-point. This is particularly true of Edge intelligent agents communicating with each other and with people. Local chat AI with RAG (Retrieval-Augmented Generation) will produce similar multi-point to multi-point traffic. There may also be a tendency for groups to share a specialized AI processor, such as a Mac Studio Ultra. For office-based work groups, this will primarily produce LAN traffic. But, for remote workers, this will produce traffic that appears to be multi-point.

A significant rise in Edge AI will substantially change traffic patterns. Some see Edge eventually displacing Data Center AI. Others suggest that the rise in demand for AI services will be such that Data Center AI will still be very active. That there will just be a change in the proportion of traffic from each one. Either way, there will be very significant changes in the traffic pattern.

San Francisco Problem Conundrum

When the pandemic hit the San Francisco Bay Area, practically no one traveled to the financial district. The hub and spoke network had no traffic. There was no user revenue coming in. System managers felt that with government assistance and reserves on hand, they could weather the storm until the pandemic waned and things returned to 'normal'. When the Pandemic waned, much of the work stayed at the Edge. Work had evolved into a hybrid remote / office pattern. There was now some traffic and user revenue going to the financial district, but not enough to support the system. Also, users wanted more multi-point to multi-point services that the current hub and spoke system was not configured for. As this is written, the public transportation system operators are struggling with how to respond to the new traffic pattern.

The risk is that WAN network operators, in responding to the current and expected rise in demand for hub and spoke networks, will find themselves in the same position as San Francisco public transportation providers when Edge AI grows substantially.

The Need for Hybrid Network Architectures

Rather than get caught in the San Francisco conundrum, it seems prudent to design the AI networks as hybrid networks from the beginning. Don't wait till the traffic patterns change and end up in crisis trying to respond. This can be thought of as using a portfolio management approach to lower risk. That is, build for existing traffic based on an understanding of likely

near-term traffic growth patterns. But also provide infrastructure for the expansion of multi-point to multiple-point networks.

While building and operating in this hybrid mode, it is important to constantly study the development of AI to inform the projection of near-term traffic growth. Ongoing study of AI technology and adoption evolution needs to be built into traffic models. But projections may always be imperfect. So, it is important to design networks to be able to grow and morph as AI traffic patterns evolve.

Conclusion

Today, our networks are being transformed by evolving AI traffic patterns. This transformation is not a single event. Rather, to be successful, network leaders have to anticipate and plan for a series of changes in traffic patterns triggered by developments in GenAI and supporting hardware.